COALFIRE®
DivisionHex™
fear no future

Collapse of Enterprise Security Timelines

The Implications of Offensive Advancements and What Organizations Must Do Now to Survive

By: John Hendley, Vice President Offensive Security

Contributing Editors: Danny Akacki, Yannick Bedard, Jacob Carlson, Justin Podzunas, Michael Raibick, Alex Reid, Neil Wyler

A DivisionHex Whitepaper | Coalfire

Date: April 2026

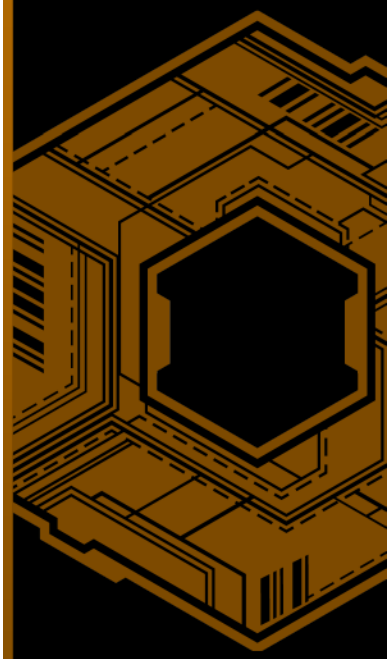

Table of Contents

Executive Summary

Offensive security abruptly shifted forward in April 2026. Leaders in enterprises and governments need to act now.

Anthropic announced that Claude Mythos Preview, a frontier AI model that has yet to be publicly released, autonomously discovered thousands of previously unknown zero-day vulnerabilities across every major operating system and every major web browser. Some of these flaws had been hiding in production code for over two decades. Of note, the model didn't just find the vulnerabilities but created working exploits.

In controlled testing, it completed 73%[1] of expert-level capture-the-flag challenges that no AI model had ever solved before, and it executed an average of 24 of 32 steps in a simulated corporate network attack chain (though it did achieve all 32 steps a third of the time). This prompted the U.S. Treasury Secretary and Federal Reserve Chair to convene an emergency meeting[2] with the CEOs of the nation's largest financial institutions.

That meeting happened because the implications are straightforward: the same capabilities that defenders are now using to find and patch vulnerabilities will be in the hands of attackers. But this isn't coming in some distant future. Within six to 18 months, openly available AI models will reach parity on vulnerability discovery. A few months after that, autonomous exploitation will follow. OpenAI released its own cyber-specialized model, GPT-5.4-Cyber, merely one week after the Mythos announcement. The capability curve is accelerating.

Here is the critical insight that most of the public commentary has missed: AI is not changing attacker tactics; rather, it's compressing the timeline. The kill chain that a skilled human adversary executes over days or weeks will be executed in minutes. While the playbook is familiar[3], the speed is not.

This creates an existential problem for any organization whose detection, response, and containment capabilities are designed for human-speed threats. If your mean time to detect is measured in hours and your incident response escalation chain requires executive-level approval before significant containment actions are taken, you have built a security program for a world that no longer exists.

The time for action is now. While these capabilities have not yet proliferated to the point of widespread operationalization by threat actors, the remaining window is measured in months, not years.

This paper provides a concrete framework for action across four strategic pillars:

- Close visibility gaps,
- Accelerate vulnerability management,
- Harden architecture to be more hostile to threats, and
- Build the foundation for autonomous defensive capabilities.

Each section contains specific, actionable recommendations, grounded in what DivisionHex's team of hackers and defenders observe every day inside the environments of the world's largest enterprises. Executing on this roadmap will be time consuming, and in some cases, costly to implement. However, organizations that execute on this framework will be better positioned to survive what comes next.

[1] https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities
[2] https://www.bloomberg.com/news/articles/2026-04-10/anthropic-model-scare-sparks-urgent-bessent-powell-warning-to-bank-ceos
[3] For reference, our own team have used publicly available models to identify, exploit, and write exploitation tools used on client engagements with great success in recent months.

Scenario: GLASSBREAK

Let's level set the problems against a plausible example scenario set 12–18 months from now. Every element is grounded in capabilities that exist today or are on a documented development trajectory. Most of what we describe are attack capabilities our teams are either already deploying in client environments or trialing internally.

The Threat: A new threat group emerges, calling itself "s0ma." In structure, it looks like dozens of other ransomware and extortion-as-a-service operations that have professionalized over the past five years. Dedicated developers maintain the tooling. Initial access brokers who supply entry points into target environments. Affiliates run the operations. Negotiators manage the extortion.

What makes s0ma different from its predecessors is a single addition to the operational stack: an AI-augmented attack automation platform built on fine-tuned open-source models. No frontier models or stolen API keys. Simply openly available weights, purpose-built tooling, and a team that understands how to operationalize them. This is the same trajectory the ransomware ecosystem has followed for many years. First, from manual encryption campaigns, then to affiliate programs, to multi-extortion enterprises with specialized roles. Only now, the next phase of professionalization is automation of the attack chain itself.

The Target: TransRouting Financial is a fictitious multinational financial services institution. 40,000 employees across 30 countries. Regulated by multiple jurisdictions. They have a mature security program: a 200-person security organization, a 24/7 SOC, endpoint detection and response deployed across the enterprise, a vulnerability management program, annual penetration testing, and a board that receives quarterly cybersecurity briefings. By most industry benchmarks, TransRouting is above average.

However, TransRouting Financial also has the same gaps that DivisionHex finds in virtually every engagement against organizations of this profile: EDR agents that were deployed to 90% of endpoints (the remaining 10% are legacy systems, acquisition integrations, and "temporary" exceptions that somehow became permanent), logging that is enabled on production systems but inconsistent across development and staging environments, network segmentation that exists in architecture diagrams but has eroded through years of firewall rule exceptions, and a VPN concentrator that was last patched four months ago because the maintenance window kept getting deferred.

T+00:00 — Initial Access

s0ma's agentic platform identifies a zero-day vulnerability in a legacy internet-facing IIS server. The model identifies the flaw, develops a working exploit, and establishes a foothold on the server, granting SYSTEM-level access. The first indicator of compromise is process execution on a host that serves static webpages and was deprioritized for monitoring.

T+00:01 – 00:08 — Reconnaissance and Credential Harvesting

From the compromised IIS server, the automated platform performs internal network reconnaissance at machine speed. It maps the Active Directory environment, identifies trust relationships, enumerates service accounts, and locates systems with cached credentials. The finely tuned agents complete the tasks in minutes, with zero reliance on human operators.

The platform identifies a service account with excessive privileges that has an active session on the host. Credential guard was never enabled, and the cleartext credentials are read from memory. The misconfiguration was previously identified in penetration reports from the last three years but was never prioritized for remediation since the host wasn't a "priority system."

T+00:09 – 00:25 — Privilege Escalation and Lateral Movement

Using the harvested service account credentials, the platform pivots towards escalating privileges to domain administrator (DA) within the Windows AD environment. It moves laterally to file servers, database systems, and backup infrastructure and uses the same misconfigurations that human attackers have exploited for years: kerberoasting against weak service account passwords, authentication coercion against hosts that were never appropriately hardened, and abuse of Group Policy Objects to push execution to additional systems.

The platform traverses network segments that exist on paper but not in practice. Firewall rules accumulated over years of exceptions allow traffic that the original architecture never intended. The malicious agents don't need to know the history. They simply probe for and exploit the lowest hanging fruit.

DA-level access is achieved, and the domain controller's NTDS.dit file is extracted, granting full access to usernames and NTLM password hashes within the main AD domain[4].

T+00:26 – 00:50 — Data Staging and Exfiltration

Leveraging privileged accounts, the agents identify and stage sensitive data: customer records, transaction histories, proprietary trading algorithms, internal communications referencing regulatory matters. It compresses and encrypts the data, then performs exfiltration through outbound HTTPS connections that blend with legitimate cloud service traffic. Data loss prevention tools are configured to

[4]Through this stage, the attack mirrors DivisionHex's own R&D efforts. Our automated tooling (using zero frontier models) achieves full AD domain forest compromise in approximately eight minutes in realistic lab environments. In production, the complexity increases, but the fundamental mechanics are identical, and privileged access is still measured in minutes. The gaps agents exploit are the same gaps our human operators exploit. The difference is speed.

inspect common exfiltration channels, but the AI-selected exfiltration path uses a sanctioned cloud service endpoint, effectively hiding in plain sight.

T+00:51 – 00:60 — Ransomware Deployment

With data already exfiltrated, the platform deploys ransomware across domain-joined systems simultaneously. Backup systems, which the platform identified and compromised during lateral movement, are encrypted or wiped first. The ransomware detonates across thousands of endpoints within minutes, taking down core revenue-producing IT systems.

T+00:61 — The Clock Has Already Run Out

TransRouting's SOC receives its first high-fidelity alert at approximately the 40-minute mark, triggered by anomalous Kerberos authentication patterns. The Tier 1 analyst reviews the alert, determines it requires escalation, and pages the on-call incident commander. By the time the incident commander joins a bridge call and authorizes containment actions, the ransomware has detonated. The IR playbook that was designed for incidents that unfold over hours never had a chance to execute.

The extortion demand arrives via TransRouting's public-facing contact form: pay $40 million or the exfiltrated data, including customer PII and regulatory correspondence, will be published in 72 hours. s0ma's negotiation team, perhaps ironically still a human function, is standing by.

What This Scenario Tells Us

The Glassbreak scenario did not require any capability that is beyond what current technology can deliver or what credible projections place within an 18-month horizon. Every tactic in the kill chain, from perimeter infrastructure exploitation, credential harvesting, Kerberoasting, lateral movement through poorly segmented networks, to exfiltration through sanctioned cloud services, is in active use by human adversaries today. The only variable that changed was speed.

And the only reason speed was decisive is because TransRouting's environment had the same gaps that nearly every enterprise has: incomplete endpoint coverage, inconsistent logging, segmentation that eroded over time, stale service accounts, and an incident response process designed for a slower adversary.

With that scenario as context, let's examine the threat landscape shift and what practitioners need to do about it.

Analysis In-Depth

Before diving into what to do, it is worth being precise about what is actually happening, because the industry conversation around Mythos has generated a significant amount of noise.

First, AI models are not inventing new attack techniques. They are executing well known techniques at a pace that collapses the defender's decision loop to irrelevancy. Full reconnaissance that once took a skilled operator a couple of hours now takes a couple of minutes. Exploit development that takes days or weeks becomes hours. Full kill chain execution that takes weeks becomes an afternoon, and unfortunately soon, it will be less than that.

Simultaneously, the barrier to entry is collapsing. Research from AISLE demonstrated[5] that 100% of AI models tested (including one with merely 3.6 billion parameters) successfully identified the same critical FreeBSD vulnerability that Anthropic showcased as a headline Mythos achievement. The vulnerability discovery capability is already broadly accessible. But the frontier moves fast. What is restricted to Mythos today will be available in open-source models that anyone can run locally within six to 18 months, according to estimates from both Anthropic's[6] own red team and Wiz[7].

The parallel to ransomware's evolution over the past seven years is instructive. In 2019, ransomware operations were driving a largely manual process: it took over 2 months on average for the TrickBot-to-Ryuk attack path to finish[8]. By 2025, the ecosystem had professionalized into an industry with specialized roles: developers maintaining malware, initial access brokers selling footholds, affiliates executing operations, negotiators managing extortion, and in some cases, dedicated PR teams running media pressure campaigns. Mandiant highlighted an operationalization milestone in their 2026 M-Trends report: they noted a 22 second window in 2025 for initial access handoff[9].

Operationalization did not require new attack techniques, simply organizational maturity, arising through a form of natural selection where only the fittest get paid. AI-augmented offensive operations will follow the same pattern, but the maturation curve will be much faster because the building blocks will be widely available.

Conversely, and this matters tremendously, what matters for defenders is that the fundamental principles of defense still apply. Organizations with comprehensive visibility, disciplined patch management, sound architecture, and practiced response capabilities will be harder targets regardless of whether the adversary is a human with a keyboard or a swarm of agents executing tasks.

What to do about it

To deal with the existential threat that these offensive advancements represent, we propose a four-pillar approach. If your organization adopts and executes these initiatives within the timeframe suggested, you will make yourself a much more difficult organization to attack. The pillars are:

1. Be maniacal about visibility,
2. Supercharge your vulnerability management program,
3. Fix your architecture, and
4. Implement autonomous defense.

[5] Stanislav Fort, AISLE https://aisle.com/blog/ai-cybersecurity-after-mythos-the-jagged-frontier
[6] https://www.nbcnews.com/tech/security/anthropic-claude-mythos-ai-hackers-cybersecurity-vulnerabilities-rcna273673
[7] Ami Luttwak, Wiz https://www.wiz.io/blog/claude-mythos
[8] John Dwyer, IBM X-Force https://www.ibm.com/think/x-force/analysis-of-ransomware
[9] https://cloud.google.com/blog/topics/threat-intelligence/m-trends-2026

Each of these pillars grow out of the others. You cannot remediate what you can't see, and as your vulnerability management program runs, you will be presented with opportunities to fix your architecture. Once you have those security fundamentals better handled, you can move to deploying autonomous defense over the top.

Pillar One: Be Maniacal About Visibility

This is the foundation of your new security program. Nothing else in this document matters if you cannot answer this simple question with confidence: what is happening across every system, endpoint, identity, and network segment in your environment right now (and can you correlate those events)?

The DivisionHex team has operated inside the environments of major enterprises; financial institutions, healthcare systems, critical infrastructure operators, Fortune 500 companies, you name it, we've hacked or secured it. I can tell you with certainty that the answer is almost never "yes, we have full visibility." The gaps are consistent and predictable. Common examples we see include:

- Endpoint Detection and Response (EDR) agents deployed to 90% or more of endpoints (less on servers), with the remainder consisting of legacy systems, acquired environments, and exceptions that were supposed to be temporary but ended up permanent due to business requirements, lack of staffing, or some combination thereof[10].
- Overreliance on EDR in general, especially given the frequent targeting of those controls by malicious actors.
- Overreliance on signature-based payload detection, especially as the cost to create custom payloads decreases significantly.
- Logging enabled on production systems but not on connected development, staging, or disaster recovery environments.
- Lack of telemetry health stream monitoring.
- Detection rules written for a threat model three years old and never re-tuned (if written at all).
- SIEM platforms ingesting data from many sources but not correlating across them effectively, often due to overlap between security and operational SIEM use cases.
- Cloud workloads spun up by development teams with no security telemetry configured.

This list is far from exhaustive. In every case, the organization believed their coverage was better than it actually was.

In the context of AI-augmented threats, these gaps are not minor deficiencies. They are the seams that malicious agents will find and exploit in minutes. If an edge device is not sending telemetry to your SOC, the initial compromise is invisible. If a development environment is not monitored, lateral movement through it is undetected. If your detection rules do not fire on the specific technique being used, you are relying on an already-overwhelmed analyst to notice something anomalous in a sea of noise. At machine speed, the window to notice shrinks to mere minutes.

Action plan

To help facilitate action in your environment, we've structured steps organizations should take around detection, both now (3 months) and in the near future (3-12 months). This list is not exhaustive, but if you follow these steps, you should be on your way to preparedness for these emerging threats.

[10] While every organization should strive for as much visibility as possible, there are many considerations that warrant a separate, more in-depth discussion. From cost of deployment to OT, IoT, and legacy systems, sometimes it's just not practical to get to 100%. That's where the importance of compensating controls from the other pillars are key.

Actions organizations should take in the next three months:

- Conduct a comprehensive audit of your endpoint detection coverage. Not a spreadsheet review, but an actual reconciliation of deployed agents against your asset inventory, with validation that agents are functioning, reporting, and on current signature and behavioral detection versions. DivisionHex has developed structured Security Tool Gap Assessment methodologies specifically for this purpose, and the most common finding is that the gap between perceived coverage and actual coverage is larger than anyone expected, so be prepared to deploy fixes post-review.

- Validate that logging and audit policies are enabled and flowing for every network segment, every identity provider, every cloud environment, and every externally facing system. Prioritize the areas where your SOC has the least visibility today; these are the areas an automated attacker will find first.

- Review your alerting and detection rules and procedures against the MITRE ATT&CK techniques most commonly used in post-compromise activity: OAuth/session token theft, cloud and hybrid identity manipulation, Kerberoasting, credential dumping attacks, lateral movement both on-premise and through cloud pivots, and exfiltration through sanctioned cloud services. If your detection engineering team has not tuned these rules in the past six months, assume they need work. Assume that the rules you tune now will be evaded over time and build a process to periodically test and recalibrate.

Actions organizations should take in the next 3-12 months:

- Invest in correlation capabilities that can connect signals across endpoint, network, identity, and cloud telemetry in near-real-time. Individual alerts are insufficient when the attack chain executes in under an hour. You need the ability to recognize a pattern across multiple data sources within minutes, not hours.

- Extend monitoring to every environment that has network connectivity to production, including development, staging, QA, and disaster recovery. These environments are routinely less monitored and less segmented, making them ideal pivot points for lateral movement.

- Enhance your detection testing program: regular, scheduled exercises in which your offensive security team or an external partner actively tests whether your detection capabilities fire against specific techniques. Do not assume your detections work. Prove it.

Metrics that matter:

You get what you measure. These are some important detection KPIs you need to monitor to identify where progress is being made (or where to course correct):

- Percentage of assets with functioning, reporting EDR agents versus total asset inventory,
- Percentage of network segments with logging flowing to your SIEM or correlation platform,
- Mean time from technique execution to alert generation, measured through regular offensive detection testing exercises

Pillar Two: Turbocharge Vulnerability Management

Pillar One is about knowing what you have. Pillar Two is about closing the gaps in what you know is broken.

Most enterprise vulnerability management programs were designed for a world in which the pace of new disclosures was relatively predictable and the time between disclosure and active exploitation was measured in weeks or months. That world does not exist anymore.

When AI models can discover and weaponize a zero-day vulnerability in hours, and when the volume of disclosed CVEs in foundational software is spiking due to AI-augmented defensive research, the traditional vulnerability management cadence of monthly scan, quarterly review, and annual penetration test is insufficient.

According to Palo Alto, in 2025, the fastest 25% of intrusions reached exfiltration in just 72 minutes, which is 4x faster than the 2024 benchmark (285 minutes)[11]. The margin between "we will patch that next cycle" and "we have been breached" was already razor thin. AI compression will eliminate it.

Actions organizations should take in the next three months:

- Prioritize remediation of every known-exploitable vulnerability in your environment. To emphasize, not every vulnerability, but every *exploitable* vulnerability. If a CVE has a weaponized exploit in the wild and you have an affected system in your environment, that is a critical path item. The CISA Known Exploited Vulnerabilities catalog is the minimum starting point. When further prioritization is required, focus on remediating exploitable vulnerabilities that are connected to technology and processes that would have the greatest impact for the organization if taken offline.
- Review your patching cadence for edge devices: VPN concentrators, load balancers, firewalls, remote access gateways, etc. These are the systems that AI-augmented attackers will target first, and they are consistently the systems with the longest patch latency because maintenance windows are difficult to schedule.
- If you don’t have a patching solution in place already, begin implementing tools that facilitate scalable patching across the different sections of your IT environment (Windows server, endpoints, Linux/Unix, etc.).

Actions organizations should take in the next 3-12 months:

- Build a process that can respond to newly disclosed vulnerabilities with emergency patching or immediate segmentation within hours, not days. This requires pre-authorized change windows, tested rollback procedures, and a decision framework that can determine whether a given vulnerability warrants emergency action without convening a committee. This will require a collaboration between security teams and “the business” deeper than most organizations have today and will likely need to be driven at the C-level, as the incentive structure to collaborate doesn’t exist today in most enterprises.
- Integrate threat intelligence into your vulnerability prioritization: does a working exploit exist, is it being actively used, and does it affect systems in my environment that are reachable from the attack paths I care about? Tenable research has shown that less than 3% of vulnerabilities are

[11] https://www.paloaltonetworks.com/resources/research/unit-42-incident-response-report

ever exploited in the wild;[12] however, the ones that are, can be weaponized shortly after disclosure.

Medium-term actions organizations should take in the next 12–36 months:

- Begin incorporating AI-assisted vulnerability scanning and code analysis into your AppSec program. The same model capabilities being used offensively will be available for defensive scanning. Organizations that build this capability now and include the internal expertise to interpret and act on AI-generated findings will have a meaningful advantage over those that wait.
- Prepare your vulnerability management program for the deluge of new vulnerabilities. Establish triage processes that can handle a significant increase in disclosure volume for foundational software without degrading your response times on the vulnerabilities that matter most to your specific environment. This is especially important for organizations that have a significant number of in-house developed applications and services.

Metrics that matter:

These are some important vulnerability KPIs you need to monitor to identify where progress is being made (or where to course correct):

- Mean time to remediate known-exploitable vulnerabilities in externally facing systems.
- Percentage of edge devices or applications patched within 72 hours[13] of critical vulnerability disclosure.
- Ratio of vulnerabilities remediated versus vulnerabilities “accepted” as risk, tracked over time.

Additional consideration for regulators: Change control processes and risk requirements in highly regulated industries can sometimes make meeting a 72-hour patch window practically impossible. It’s not just security that needs to change to address the speed of AI, it’s the regulatory and process changes that are required as well.

[12] https://coalfire.com/insights/news-and-events/press-releases/divisionhex-fires-back-at-industry-confusion-with-a-clear-roadmap-for-exposure-management

[13] This is the minimum starting point. Organizations will need to accelerate this to same day as fast as possible.

Pillar Three: Hostile Architecture

Visibility tells you what is happening. Vulnerability management fixes what is broken. Architecture determines how much damage an attacker can do once they are inside. For more than a decade now, our team has been insisting that compromise is a matter of when, not if. Architecture is your primary mechanism for buying time.

If an AI-augmented attacker can move from a compromised edge device to domain administrator in minutes, the only thing that slows them down is encountering boundaries that require additional exploitation, additional credential theft, or additional reconnaissance. Every segment they must cross, every hash they must crack, every trust boundary they must traverse is friction. Enough friction turns a 60-minute attack into a multi-hour attack. That may be the difference between a contained incident and a catastrophic breach.

The problem is that most enterprises have architecture that looks great in diagrams but has eroded significantly in practice. Our team has seen firewall rule tables accumulate numerous exceptions. Service accounts are provisioned for projects and never decommissioned. External test systems are forgotten about and unpatched. Administrative credentials are reused across environments. Network segments that were designed to isolate sensitive systems have been bridged by exception requests that became permanent. DivisionHex sees this in every environment we operate in, regardless of the organization's size or sophistication. Below are the steps organizations can take to create architecture that’s hostile to attackers, resulting in increased dwell time.

Actions organizations should take in the next three months:

- Audit your perimeter. Every externally facing service, device, and application is a potential foothold. If it does not need to be internet-facing, remove it. If it must be exposed, ensure it is patched, monitored, and segmented from internal networks wherever possible. Do not rely on your existing asset inventory. Refresh it with a heavy focus on open-source intelligence (OSINT) and discovery tooling (like attack surface management toolsets).

- Conduct a service account inventory. Identify every service account with privileged access, determine whether it is still required, and enforce credential rotation. Stale service accounts with excessive privileges are one of the most reliable pivot points in enterprise compromise. Our team exploits these in the majority of our internal network offensive engagements.

- Enforce MFA on every identity, including administrative and service accounts where technically feasible. Implement strict detection and monitoring in areas where it is not feasible.

- Eliminate password reuse across environments. During offensive security engagements, DivisionHex routinely finds privileged accounts in different domains (or even forests!) that share passwords. Unique passwords must be enforced for every account.

Actions organizations should take in the next 3-12 months:

- Implement or validate network segmentation between business-critical systems, general corporate infrastructure, development environments, and administrative planes. Test the

segmentation, not by reviewing firewall rules, but by having a human behind the keyboard actually attempt to traverse the segments and documenting where they succeed.

- Adopt a tiered administrative model that prevents domain administrator credentials from being used on standard workstations.

- Restrict lateral movement paths by:
    - limiting line of sight between systems that don't need to directly communicate,
    - limiting where accounts can authenticate to endpoints and servers, and
    - limiting network protocols that are permitted across segment boundaries.

Medium-term actions organizations should take in the next 12–36 months:

- Move toward zero-trust architectures that enforce identity verification and device health at every access decision, not just at the perimeter. Please note that you cannot "buy a zero trust" (yes, this is something we've heard a product sales rep actually say before). This is an architectural transformation. It requires rethinking how access is granted, how trust is established, and how sessions are monitored throughout their lifecycle.
- Design architecture with the assumption that any single component, including any endpoint, any server, any identity, any edge device — may be compromised at any time, and the blast radius of that compromise must be contained by design rather than by detection.

Metrics that matter:

- Number of externally facing services and devices, tracked monthly with a reduction target.
- Percentage of privileged accounts with MFA enforced.
- Number of firewall rule exceptions older than 12 months. Each one is a potential traversal path.
- Number of accounts that share the same password (this should be zero, at least for privileged accounts)[14].

[14] Perform password audits where unable to know where passwords are reused – we have developed methodologies specifically for this purpose.

Pillar Four: Autonomous Defense

The first three pillars are prerequisites and, frankly, aren't new ideas. This is where we start getting into the future. And this is by far the most contentious topic. Automated defense is still highly debated and largely untested in complex real-world environments. The caution against saying "put AI in this" is warranted. But if the adversary is operating at machine speed, the defense must eventually operate at machine speed too.

No human SOC team, regardless of skill, can detect a full kill chain executing in under 60 minutes, escalate through an approval chain, and execute containment actions in time to prevent the objective. To be clear, humans are not too slow because they lack talent. They are too slow because they are human. This does not mean removing humans from the loop. It means redesigning the loop.

The concept is defender-in-the-loop autonomous response: AI-driven security agents that continuously correlate signals across endpoint, network, identity, and cloud telemetry, identify patterns indicative of active compromise, recommend specific containment actions, and (depending on your organizational risk profile) execute those containment actions autonomously.

For higher-impact actions, like isolating a business-critical system, revoking access for a senior executive, or shutting down a network segment, the agent presents its recommendation with full business context and waits for human approval. The human makes the judgment call. The agent gives them the information to make that call in minutes rather than hours.

This requires organizations to make decisions now that they have historically deferred. Examples of these hard decision points include:

- What actions can an automated system take without human approval? Isolating a single endpoint? Blocking a suspicious IP? Revoking a compromised credential? These seem straightforward in theory, but in the real world they rapidly take on significant operational risk, and most organizations have not explicitly defined the authority boundaries.
- What business context does the agent need to make sound recommendations? An agent that recommends isolating a trading platform during market hours without understanding the financial impact is not helpful, it's dangerous. Therefore, the agent must also understand the operational significance of the systems it is protecting.
- What governance framework surrounds autonomous defensive actions? This is not a technology question. This is a board-level conversation about risk appetite, accountability, and the acceptable trade-off between speed of response and risk of false positive disruption.

The organizations that begin building the visibility infrastructure, the correlation engines, the pre-authorized response playbooks, the governance frameworks, etc. will be ready when the technology matures. The ones that wait won't survive.

There aren't timelines listed for this section, because every organization needs to fully operationalize the first three pillars as described as a prerequisite.

Addendum One: Most IR Playbooks are Already Obsolete

While not its own pillar, your IR playbook deserves its own callout because it sits at the intersection of every other initiative and is overlooked in almost every security program review.

Most enterprises (financial and other highly-regulated institutions in particular) have incident response playbooks that assume a specific tempo: the attack unfolds over hours or days, the SOC detects anomalous activity, the incident is escalated through a defined chain, legal counsel is engaged, containment actions are authorized by leadership, forensic investigation begins, regulatory notification timelines are assessed, and board communication is prepared. These playbooks represent years of refinement and regulatory alignment, and they are designed for a world where the adversary operates at human speed.

When the attack takes 60 minutes, the playbook never reaches Step 3.

Organizations need to redesign their response processes for a compressed timeline. This means pre-authorizing containment actions that currently require escalation. It also means establishing clear delegation frameworks so that a SOC analyst or an automated agent can isolate a compromised segment without waiting for a VP to join a bridge call. It means testing these compressed playbooks through tabletop exercises and live adversary simulations that explicitly model machine-speed attack scenarios.

If your incident response plan has not been stress-tested against a 60-minute attack scenario, it has not been tested against the threat you are about to face.

Addendum Two: Third-Party and Supply Chain Exposure

Everything discussed in this paper applies not only to your own environment but to every vendor, payment processor, correspondent institution, and technology partner in your ecosystem.

An AI-augmented attacker does not need to compromise your organization directly. It can identify the weakest link across your supply chain faster than any third-party risk management questionnaire can assess it. If your payment processor has an unpatched edge device, if your cloud provider has inconsistent logging, if your acquired subsidiary has never been fully integrated into your security program, those are all viable entry points into your environment.

The traditional TPRM model, with its annual questionnaire, evidence review, and risk rating, was designed for a world where supply chain compromise was an advanced, targeted operation.

When the cost and complexity of supply chain attacks drop to the point where they are economically viable for any professionalized cybercrime operation, the assessment model must evolve accordingly.

At minimum, organizations should be requiring continuous monitoring evidence (not point-in-time attestations) from critical third parties. They must also validate through their own testing that trust boundaries between their environment and third-party connections are actually enforced.

Final Thoughts: The Window is Open, but Closing

The capabilities demonstrated by Mythos are not the ceiling. They are the floor of what the next generation of AI models will deliver. Within six to 18 months, the offensive capabilities that are currently restricted to a handful of frontier labs will be available in open-source models that any motivated actor can deploy without restriction.

Ideologically motivated actors will use them to exploit known vulnerabilities at unprecedented volume.

Professionalized cybercrime syndicates will use them to compress full attack chains from days to minutes, following the same operationalization playbook the ransomware ecosystem has refined over the past five years. But organizations learned lessons from that. They understood the impacts from data theft, from business email compromise, from operational disruption. Those same impacts are likely to be repeated in the next several years, just executed at extreme speed.

The focus of this paper has been on the threats posed by individual actors and organized cybercrime syndicates. These are the groups that will most immediately operationalize AI-augmented offensive capabilities against the broadest set of targets. But it would be incomplete not to acknowledge that nation-state actors represent a wholly different tier of threat[15]. Their funding, sophistication, and strategic patience far outstrip the other two groups, and AI augmentation in their hands will enable capabilities that demand their own analysis and response framework. That said, the foundational work described here, including visibility, vulnerability management, architecture, and autonomous defense, is the prerequisite regardless of which threat actor is on the other end. You cannot defend against the most sophisticated adversaries if you have not yet solved the fundamentals.

The organizations that survive this transition will be the ones that treated this moment not as a headline to monitor but as a starting gun. They will have closed their visibility gaps, fixed what they knew was broken, and re-architected their environments to contain blast radius. And they will have begun building the autonomous defensive capabilities that will be the only answer to autonomous offense.

DivisionHex exists because we're passionate about security. We believe the best way to understand the threat is to provide threat-focused views on our clients' terms, under controlled conditions, with the explicit goal of making them harder to break. Our team operates inside the world's most complex environments every day. The gaps described in this paper are not theoretical. They are what we find, what we exploit, and what we help our clients fix.

The window to prepare is open. It will not stay open long.

[15] Further reading from Anthropic on this topic from November 2025: https://assets.anthropic.com/m/ec212e6566a0d47/original/Disrupting-the-first-reported-AI-orchestrated-cyber-espionage-campaign.pdf

About DivisionHex

DivisionHex is Coalfire's elite team of hackers, defenders, and threat hunters — built to test limits, break systems before attackers do, and redefine what it means to be secure. From adversary simulation and threat-informed penetration testing to threat hunting, exposure management, and cyber risk advisory, DivisionHex helps enterprises fear no future.

To discuss how your organization can prepare for AI-augmented threats, contact DivisionHex at coalfire.com/divisionhex.

Epilogue: The “Good” News

There is one last, subtle nuance in the overall discussion of Mythos deserving attention. We are entering a world where bad actors can autonomously collect bugs, with exploits, and compromise systems running vulnerable software before vendors know to react.

Zero-days will be the new normal, leaving administrators with no idea of the initial entry vector. This is terrifying for many and is a reasonable reaction. Importantly though, while LLMs are finding vulnerabilities, they are not yet finding vulnerabilities in specific, live targets. They cannot fabricate skeleton keys. They are not yet “Joshua,” “Setec Astronomy,” or “Wintermute.”

AI’s advantages in bug hunting are speed, relentlessness, and absence of fatigue; however, as previously stated, LLMs do not possess any skills beyond human vulnerability researchers. They can't. They can only learn the information we documented and forced them to read. This means that, for now, when attempting to find and exploit a new vulnerability in a specific target the AI must play by the same rules as the humans who are already doing this every single day.

This means that first, of course, an exploitable vulnerability must exist in the desired point of entry. That alone is a huge assumption. To start that hunt, the AI must create lab environments that model their target's exact attack surface, including hardware and software. As always, attack surface size is crucial. Many bugs are dependent on operational environment, i.e. their presence is a function of specific configurations or certain system and network loads that the lab must simulate.

Some bugs only exist during a confluence of highly specific states (e.g., mutex wait state or file descriptor ownership)[16]. This means that after the bug is isolated and an exploit written, the attacker may need to force the target into the exact state required to engender the vulnerability. And then, finally, the exploit must work, which any hacker will tell you requires more than good aim.

We are still a long way away from a world where an AI can set its sights on an arbitrary environment and guarantee compromise. While that’s a pretty thin silver lining, this is the "good" news: we still have time to prepare for a world where that will happen. Everything else in this document is training for “The Big One,” that moment where that capability becomes real.

[16] As an example, our team discovered a vulnerability that would pop a root shell, but only when a specific network printer was defined.